# 医学影像与自然语言处理多模态探索研究

龚宇新[1] 向 菲[1] 应 葵[2]

([1] 华中科技大学同济医学院医药卫生管理学院 武汉 430000 [2] 清华大学工程物理系 北京 100084)

〔摘要〕 **目的/意义** 实现医学影像报告的自动生成对减轻放射科医生工作负担、促进临床工作流程标准化具有重要意义。**方法/过程** 重点查找近几年公开源代码的胸部报告生成模型，开发一种基于 CDGPT 2 模型的医学影像报告自动生成方法。**结果/结论** 大参数量的语言模型在报告生成方面的优势仍有待挖掘，对模型的解码器输入进行修改后生成报告的质量不高。未来研究可采用大型数据集并结合更多临床信息来提高模型性能。

〔关键词〕 胸片；多模态；报告自动生成；注意力机制；自然语言处理

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.01.006

**An Exploratory Study of Multimodality in Medical Imaging and Natural Language Processing**

*GONG Yuxin[1], XIANG Fei[1], YING Kui[2]*

*[1] School of Medicine and Health Management, Tongji Medical College of Huazhong University of Science & Technology, Wuhan 430000, China; [2] Department of Engineering Physics, Tsinghua University, Beijing 100084, China*

〔**Abstract**〕 **Purpose/Significance** Achieving automatic generation of medical imaging reports is important for reducing the workload of radiologists and promoting the standardization of clinical workflow. **Method/Process** Focusing on finding the chest report generation models with open source code in recent years, the paper develops an automatic medical image report generation method based on the CDGPT2 model. **Result/Conclusion** The advantages of the model in report generation are still to be explored, the quality of reports generated after modifications to the decoder inputs of the model is not high. Future research could improve the performance of the model by using large datasets and incorporating more clinical information.

〔**Keywords**〕 chest radiograph; multimodal; automatic report generation; attention mechanism; natural language processing

## 1 引言

医学影像是疾病诊断的重要依据，对医学影像的解释主要由放射科医生完成。但日益增长的阅片需求给放射科医生造成较重的工作负荷。近年来，随着人工智能的快速发展，基于图像与文字的多模态研究受到关注。目前，以自然语言处理技术为主的报告生成是医学影像报告生成的主流方法，这种方法基于编码器-解码器框架[1]，解码器最初主要采用循环神经网络（recurrent neural network，RNN）和长短期记忆网络（long short-term memory，LSTM）。2017 年 Transformer 模型被提出且在诸多自然语言处理任务中表现优越，许多以其为基础的预训练语言模型被相继提出，因此 Transformer 模型及其变体逐渐替代 RNN、LSTM。目前医学影像报告

〔修回日期〕 2023-09-12

〔作者简介〕 龚宇新，硕士研究生；通信作者：应葵，博士，副教授。

生成模型的解码器主要基于单个预训练语言模型，本文拟通过在统一编码器条件下，针对胸片检查需求量大以及相应数据集较丰富的现状，对比不同预训练语言模型在胸片报告生成方面的性能。

## 2 相关工作

早期胸部影像报告的自动生成模型普遍采用自然图像描述领域的卷积神经网络 - 循环神经网络（convolutional neural network - recurrent neural network，CNN - RNN）架构，但是生成报告较短。之后对 RNN 解码器部分进行改进，如 Krause J 等[2]使用包含两个 LSTM（一个句子级，一个单词级）的多级 RNN，在生成段落方面效果有所提升，不足之处是会导致句子重复。为了提高生成报告的准确度，注意力机制也被应用于胸部报告自动生成任务中。Wang X 等[3]提出一个文本图像嵌入网络（TieNet），将多层注意力模型融入端到端的 CNN - RNN 框架中来抽取并强化重要的报告文本表示和胸部影像图像表示，用于分类和报告生成。Jing B 等[4]提出一种同时关注视觉特征和语义特征的联合注意力机制，其中语义特征基于胸部影像的疾病标签得到，不仅可以定位疾病在影像中的位置，而且使解码器在解码过程中更多地关注有意义的信息。在 Transformer 模型被提出后，Transformer 及其变体也被广泛应用于报告生成任务中。如 Chen Z 等[5]通过在 Transformer 的解码器中引入关系记忆模块，使模型在建模全局信息的同时更好地刻画影像报告中的局部结构，提升生成胸部影像报告的质量。Wu X 等[6]提出基于对比学习的多模态递归模型，通过融合视觉特征和语义特征生成胸部报告的“印象”和“发现”。有多种方法可实现报告自动生成，如基于报告生成模型采用强化学习[7]、知识图谱[8]或基于模板[9]等方法。

## 3 方法

### 3.1 胸部报告自动生成模型

3.1.1 公开源代码的相关模型 以近 3 年和使用 Transformer 或其变体作为语言模型为筛选条件，共找到 4 个公开源代码的报告生成模型，见表 1。

**表 1 公开源代码的胸部报告生成模型**

| 模型名称 | 时间（年） | 视觉模型 | 语言模型 | 模型框架 | 数据集 |
|---|---|---|---|---|---|
| R2Gen[5] | 2020 | VGG/ResNet | Transformer 基础上引入关系记忆机制 | PyTorch | IU - XRay & MIMIC - CXR |
| RATCHET[10] | 2021 | DenseNet - 121 | Transformer 的解码器部分 | TensorFlow | MIMIC - CXR |
| CDGPT 2[11] | 2021 | CheXnet | distilGPT 2 | TensorFlow | IU - XRay |
| XRG[12] | 2021 | DenseNet - 121 | Transformer 的解码器部分 | PyTorch | IU - XRay & MIMIC - CXR |

3.1.2 研究模型选择 上述 4 个模型均可成功搭建运行，经过分析各模型的模块，最后选择 CDGPT 2 模型进行研究，主要原因有以下两点。一是图像的处理及训练对计算机的算力要求较高，受制于设备资源及性能，选择相对较小的数据集 IU - XRay（7 470 张胸部影像）而非 MIMIC - CXR 数据集（371 920 张胸部影像）进行训练及测试，同时也便于结果对比分析。二是 CDGPT 2 模型的解码器部分采用语言模型 distilGPT 2，其封装性好、代码结构分明，后期在统一视觉模型的条件下，便于替换其他预训练语言模型如（GPT 2），以对比不同语言模型在报告生成方面的性能。

3.1.3 实验环境配置与模型参数设置 本文所涉及实验的相关配置情况，见表 2；模型参数，见表 3。

**表 2 实验环境配置**

| 设备 | 资源名称 | 配置情况 |
|---|---|---|
| 硬件 | GPU | NVIDIA GeForce RTX 3090 |
| | CPU | AMD EPYC 7302（8 核） |
| 软件 | 深度学习框架 | TensorFlow 2. 5. 0 |
| | 集成开发环境（IDE） | PyCharm 2021. 3. 1 |
| | 编程语言 | Python 3. 8 |

表 3 模型参数设置

| 参数名称 | 参数设置 |
|---|---|
| Vocabulary_Size | 1 001 |
| Num_Epochs | 100 |
| Optimizer | Adam |
| Learning_Rate | $1 \times 10^{-4}$ |

## 3.2 CDGPT 2 模型

3.2.1 模型结构 CDGPT 2 模型结构包括 3 部分：视觉特征、语义特征和解码器，见图 1。

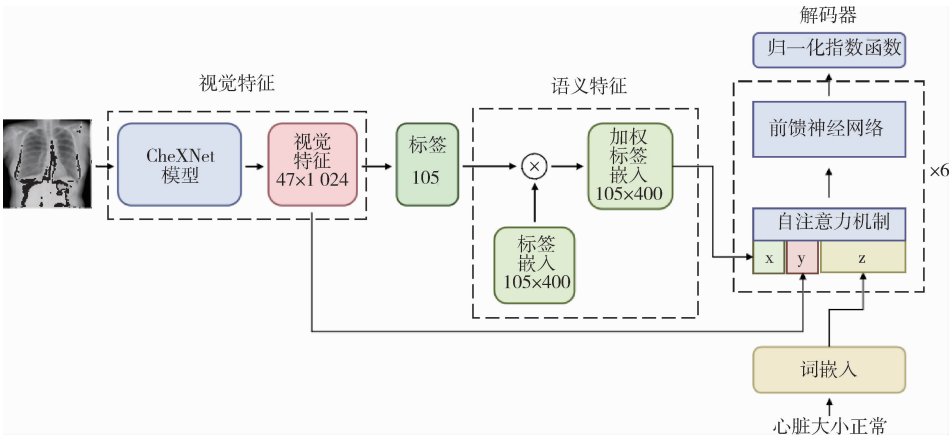


图 1 CDGPT 2 模型

视觉特征部分，输入一张胸部影像通过视觉模型得到相应的视觉特征和 105 个标签预测分数。视觉模型以吴恩达团队[13]提出的 CheXNet 模型为基础。CheXNet 模型对 14 种常见的肺部疾病进行检测及定位，但这 14 种标签不足以提供丰富多样的语义特征，因此对其进行微调。通过删除最后一层网络然后添加一层新的包含 105 个节点的网络，使其能够输出 IU - XRay 数据集中常见的 105 种手工标注标签的预测分数，每个标签的独立置信度分数在 0 到 1 之间。

$$L(b) = -\sum_{t=1}^{T}\sum_{i=1}^{N} y_{i_t} \cdot \log \hat{y}_{i_t} + (1 - y_{i_t}) \cdot \log(1 - \hat{y}_{i_t}) \tag{1}$$

式中标签的预测被视为二元交叉熵损失函数的多标签分类问题。其中，$y$ 代表真实标签，$\hat{y}$代表网络的输出值，$L$（$b$）代表第 $b$ 个批次（batch）的损失值，$T$ 代表标签的数量，$N$ 代表批次的大小（batch size）。

语义特征部分，标签嵌入来源于一个预训练 Word2Vec Embedding[14]，该预训练在大量生物医学文本中训练得到。当标签中的词多于一个时，则该标签的词嵌入（embeddding）就等于各词对应的词嵌入相加后除以词的个数，即取平均值。将 105 个标签的预测分数与标签嵌入作哈达玛积，得到加权标签嵌入一个大小为 105 ×400 的矩阵。

解码器部分，解码器以 distilGPT 2 模型为基础，在模型的输入及自注意力机制中进行了一些改动。首先是输入的改变，一般语言模型仅将词嵌入 $Z$ 作为输入，但在 CDGPT 2 模型的解码器中，增加了两个额外的输入，分别为语义特征 $X$ 和视觉特征 $Y$。其次是自注意力机制的计算方式发生了改变，由于解码器输入的增加，受限的自注意力机制（conditioned self attention，CSA）计算方式如下。查询向量 $Q$ 未发生变化，但键向量 $K$ 和值向量 $V$ 均加上了语义特征和视觉特征的信息。

$$\mathrm{CSA}(X,Y,Z) = \mathrm{softmax}\left((ZW_Q)\begin{bmatrix} XU_K \\ YH_K \\ ZW_K \end{bmatrix}^T\right)\begin{bmatrix} XU_V \\ YH_V \\ ZW_V \end{bmatrix} \tag{2}$$

式中，$W_Q$、$W_K$、$W_V$ 分别代表词嵌入 $Z$ 的查询向量、键向量和值向量的权重。同理，$U_K$、$U_V$ 分别代表语义特征 $X$ 的键向量和值向量的权重，$H_K$、$H_V$ 分别代表视觉特征 $Y$ 的键向量和值向量的权重，softmax 为激活函数。

3.2.2 替换语言模型 distilGPT 2 模型是 GPT 2 模型的压缩版本，同时也是 GPT 2 模型最小的版本，二者对比，见表 4。distilGPT 2 比 GPT 2 小 45%，训练速度快两倍，但 distilGPT 2 模型在一些质量基准上的得分要低一些。因此在成功搭建 CDGPT 2 模型的基础上，希望进一步对比 distilGPT 2 和 GPT 2 在报告生成方面的效果及训练速度。

**表 4 distilGPT 2 与 GPT 2 对比**

| 模型名称 | 层数 | 隐层维度 | 参数量（亿） |
|---|---|---|---|
| distilGPT 2 | 6 | 768 | 8.2 |
| GPT 2 | 48 | 1 600 | 15 |

3.2.3 修改解码器输入 在 CDGPT 2 模型中，语义特征 $X$ 和视觉特征 $Y$ 是在进入自注意力层后进行拼接处理的，但是这样不易替换其他类型的语言模型。一种简单易行的思路是在进入自注意力层之前将语义特征 $X$ 和视觉特征 $Y$ 拼接在词嵌入 $Z$ 中，这样就不必关注其他语言模型复杂的内部结构。这里将改动后的模型称为 Concat_CDGPT 2。

Concat_CDGPT 2 模型的具体实现细节如下，保留视觉特征和语义特征的处理不变，改动解码器的输入，即在进入自注意力层前对 $X$、$Y$ 与 $Z$ 的最后一维进行统一，然后对三者进行拼接，得到 $Z'$。此时连接的自注意力（concat self attention，Concat-SA）的计算方式如下：

$$\mathrm{ConcatSA}(Z') = \mathrm{softmax}((Z'W_Q)(Z'W_K)^T)(Z'W_V) \tag{3}$$

其中，$W_Q$、$W_K$、$W_V$ 分别代表 $Z'$ 的查询向量、键向量和值向量的权重。

## 4 实验结果与分析

### 4.1 数据集与评价指标

本文选用 IU - XRay 数据集对模型进行训练及测试。IU - XRay 数据集是一个公开的胸部影像数据集，包含 7 470 张正面和侧面胸部影像，以及对应的 3 995 份报告。在具体实验中，随机挑选 500 张胸部影像作为测试集，余下的全部作为训练集。

在对生成文本的评价中，常用的自然语言评价指标主要有 4 种，分别为 BLEU - N[15]、ROUGE[16]、METEOR[17]、CIDEr[18]，见表 5。这些指标常用于衡量生成文本与参考文本之间的相关性，一定程度上可以反映生成文本的可读性和流畅度。其数值均为越大越好。

**表 5 自然语言评价指标总结**

| 指标名称 | 针对任务 | 特点 |
|---|---|---|
| BLEU | 机器翻译 | 关注准确率 |
| ROUGE | 文本摘要 | 关注召回率 |
| METEOR | 机器翻译 | 综合考虑准确率和召回率，考虑同义词 |
| CIDEr | 图像描述 | 对不同 $n$ 元组赋予不同权重，关注重点信息 |

### 4.2 各模型结果分析

CDGPT 2 表示使用 distilGPT 2 作为语言模型，CDGPT 2（GPT 2）表示使用 GPT 2 作为语言模型，Concat_CDGPT 2 是在 CDGPT 2 基础上对解码器的输入进行了修改。

4.2.1 评价指标分析 （1）CDGPT 2 与 CDGPT 2（GPT 2）评价指标分析。CDGPT 2（GPT 2）的分数均低于 CDGPT 2，其中 BLEU 系列指标分数的差距较为明显，这说明在 n - gram 的重合度方面 CDGPT 2（GPT 2）模型的效果不及 CDGPT 2。在 CIDEr 指标中，二者也存在一定差距，这意味着 CDGPT 2 生成的报告在疾病关键信息方面的提取效果更佳，见表 6。综合各评价指标可知，CDGPT 2 的效果优于 CDGPT 2（GPT 2）。但是考虑到在模型训练中使用的 IU - XRay 数据集（7 470 张胸片）相对较小，可能使 distilGPT 2 这种轻量级语言模型占优势。而 GPT 2 的参数比 distilGPT 2 多将近两倍，网络层数及隐层维度也更多，这意味着需要更多的数据来训练，如使用 MIMIC - CXR 数据集（371 920 张胸片），才有可能体现出 GPT 2 模型的优势。（2）CDGPT 2 与 Concat_CDGPT 2 评价指标分析。Concat_CDGPT 2 模型的各指标分数均低于 CDGPT 2 模型且差距较大，尤其 CIDEr 指标的差距最大。这说明在进入自注意力层前将语义特征、视觉特征与词嵌入进行拼接的思路不可行，见表 6。通过对自注意力层计算方式的对比

及分析，Concat_CDGPT 2 模型效果较差有以下两种原因。一是查询向量 $Q$ 发生了改变。在 CDGPT 2 模型中，查询向量 $Q$ 由输入的词嵌入 $Z$ 与相应的矩阵 $W_Q$ 相乘得到。但在 Concat_CDGPT 2 中，在词嵌入 $Z$ 中拼接加入了语义特征 $X$ 和视觉特征 $Y$ 得到 $Z'$，意味着之后的查询向量也发生了改变。二是模型参数量发生了改变。在 CDGPT 2 模型中，语义特征 $X$、视觉特征 $Y$ 和词嵌入 $Z$ 均有各自的键向量及值向量，之后进行拼接得到最后的 $K$ 和 $V$。但在 Concat_CDGPT 2 模型中，键向量 $K$ 和值向量 $V$ 的计算简化，使模型的参数量变少，从而可能对模型的效果产生影响。

**表 6 各模型评价指标分数对比**

| 模型 | BLEU－1 | BLEU－2 | BLEU－3 | BLEU－4 | METEOR | ROUGE | CIDEr |
|---|---|---|---|---|---|---|---|
| CDGPT 2 | 0. 388 | 0. 247 | 0. 170 | 0. 121 | 0. 167 | 0. 304 | 0. 327 |
| CDGPT 2（GPT 2） | 0. 351 | 0. 214 | 0. 143 | 0. 098 | 0. 156 | 0. 290 | 0. 253 |
| Concat_CDGPT 2 | 0. 186 | 0. 105 | 0. 064 | 0. 037 | 0. 108 | 0. 191 | 0. 060 |

4.2.2 生成报告分析 目前，对胸部报告生成的评价没有精确标准，自然语言评价指标只能在一定程度上反映生成报告的质量。因此对各模型所生成的报告进行定性分析，相关胸部影像的真实报告和预测报告，见图 2，其中带下划线的文本表示模型可以检测到的异常并且与真实报告具有相似的描述，蓝色文本表示只在预测报告中出现而未在真实报告中出现的异常信息。真实报告中描述左肺有 5 毫米的钙化性肉芽肿，CDGPT 2 生成的报告准确地检测到了该病变并且与真实报告的描述很接近。而 CDGPT 2（GPT 2）生成的报告中虽然有提及肉芽肿病变，但是缺少位置信息，而且还出现右侧基底结节钙化和组织胞浆菌病的描述，但是真实报告中并没有这些异常信息。模型 Concat_CDGPT 2 生成的报告不断重复“肺部清晰、心脏大小正常”的语句，同时缺失对病变关键信息的描述，与 CIDEr 指标的低分数相对应，并且出现假阳性。

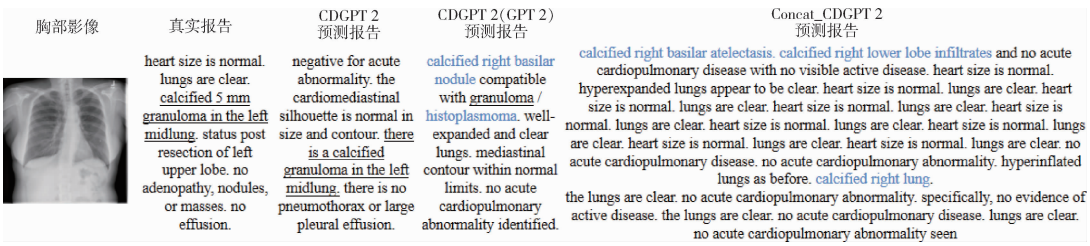


**图 2 各模型预测报告对比**

## 5 结语

针对胸部报告的自动生成，调研相关工作进展及公开源代码的模型，在对多个生成模型进行成功搭建运行后选择 CDGPT 2 模型进行研究，使用 IU－Xray 数据集的 7 470 张胸部影像对模型进行训练及测试。受限于数据集大小和计算资源，GPT 2 这种大参数量（15 亿）的模型在报告生成方面的优势还有待挖掘。另外，模型解码器的输入以及模型参数量的改变对模型有很大的影响。未来研究可进一步扩大数据集规模，结合更多临床信息，如将影像数据与电子病历数据相结合来提升模型性能。

**利益声明**：所有作者均声明不存在利益冲突。

## 参考文献

1 梅周俊森，孙水发，李小龙．基于深度学习的医学影像报告自动生成研究综述［J］．长江信息通信，2023，36（5）：21－24.

2 KRAUSE J, JOHNSON J, KRISHNA R, et al. A hierarchical approach for generating descriptive image paragraphs [C]. Honolulu: 30th IEEE Conference on Computer Vision and Pattern Recognition, 2017.

3 WANG X, PENG Y, LU L, et al. TieNet: text - image embedding network for common thorax disease classification and reporting in chest X - Rays [C]. Salt Lake City: IEEE Conference on Computer Vision and Pattern Recognition, 2018.

4 JING B, XIE P, XING E. On the automatic generation of medical imaging reports [C]. Melbourne: 56th Annual Meeting of the Association for Computational Linguistics, 2018.

5 CHEN Z, SONG Y, CHANG T H, et al. Generating radiology reports via memory - driven transformer [C]. Online: Conference on Empirical Methods in Natural Language Processing, 2020.

6 WU X, LI J, WANG J, et al. Multimodal contrastive learning for radiology report generation [J]. Journal of ambient intelligence and humanized computing, 2022, 14 (8): 11185 - 11194.

7 HOU D, ZHAO Z, LIU Y, et al. Automatic report generation for chest X - Ray images via adversarial reinforcement learning [J]. IEEE access, 2021 (9): 21236 - 21250.

8 ZHANG Y, WANG X, XU Z, et al. When radiology report generation meets knowledge graph [C]. New York: 32nd Innovative Applications of Artificial Intelligence Conference, 2020.

9 WANG X, ZHANG Y, GUO Z, et al. TMRGM: a template - based multi - attention model for X - Ray imaging report generation [J]. Journal of artificial intelligence for medical sciences, 2021, 2 (1 - 2): 21 - 32.

10 HOU B, KAISSIS G, SUMMERS R, et al. RATCHET: medical transformer for chest X - Ray diagnosis and reporting [C]. Online: International Conference on Medical Image Computing and Computer Assisted Intervention, 2021.

11 ALFARGHALY O, KHALED R, ELKORANY A, et al. Automated radiology report generation using conditioned transformers [J]. Informatics in medicine unlocked, 2021 (24): 100557.

12 NGUYEN H, NIE D, BADAMDORJ T, et al. Automated generation of accurate & fluent medical X - Ray reports [C]. Punta Cana: Conference on Empirical Methods in Natural Language Processing, 2021.

13 RAJPURKAR P, IRVIN J, ZHU K, et al. CheXNet: radiologist - level pneumonia detection on chest X - Rays with deep learning [EB/OL]. [2023 - 09 - 12]. https://arxiv.org/abs/1711.05225.

14 MCDONALD R, BROKOS G I, ANDROUTSOPOULOS I. Deep relevance ranking using enhanced document - query interactions [EB/OL]. [2023 - 09 - 12]. https://arxiv.org/abs/1809.01682.

15 PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]. Philadelphia: 40th Annual Meeting of the Association for Computational Linguistics, 2002.

16 LIN C Y. Rouge: a package for automatic evaluation of summaries [C]. Barcelona: Text Summarization Branches Out, 2004.

17 BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments [C]. Ann Arbor: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005.

18 VEDANTAM R, ZITNICK C L, PARIKH D. Cider: consensus - based image description evaluation [C]. Boston: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

(上接第 32 页)

12 郑承宇, 王新, 王婷, 等. 基于 ALBERT - TextCNN 模型的多标签医疗文本分类方法 [J]. 山东大学学报 (理学版), 2022, 57 (4): 21 - 29.

13 ZHANG X, ZHANG Y, ZHANG Q, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods [EB/OL]. [2023 - 09 - 27]. http://dx.doi.org/10.1016/j.ijmedinf.2019.103985.

14 罗玮. 患者投诉中安全事件的自动识别研究 [D]. 武汉: 华中科技大学, 2019.

15 唐仕肖, 李秀云, 李文娟. NGT 与德尔菲法应用于医院护理不良事件管理系统的分析与研究 [J]. 智慧健康, 2022, 8 (28): 189 - 193.

16 朱未, 胡少科. 基于 Reason 模型的医疗器械不良事件的影响因素研究 [J]. 生物骨科材料与临床研究, 2018, 15 (3): 77 - 80.